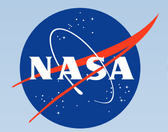
An introduction to phylogenomics

Inferring evolutionary relationships between organisms

Mike Lee
@AstrobioMike
microbialomics.org



GeneLab
Open Science for Life in Space

Blue Marble Space
Institute of Science

Phylogenetics and phylogenomics

Single-copy core genes

Which genes should we use?

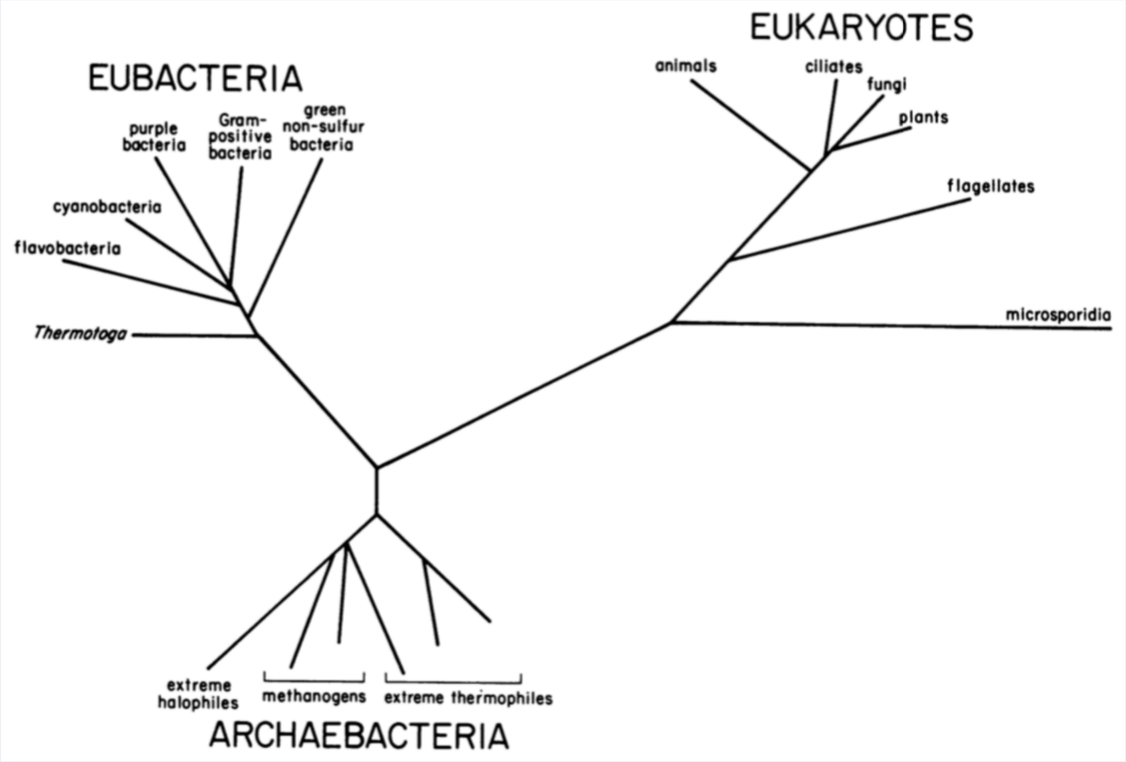
Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

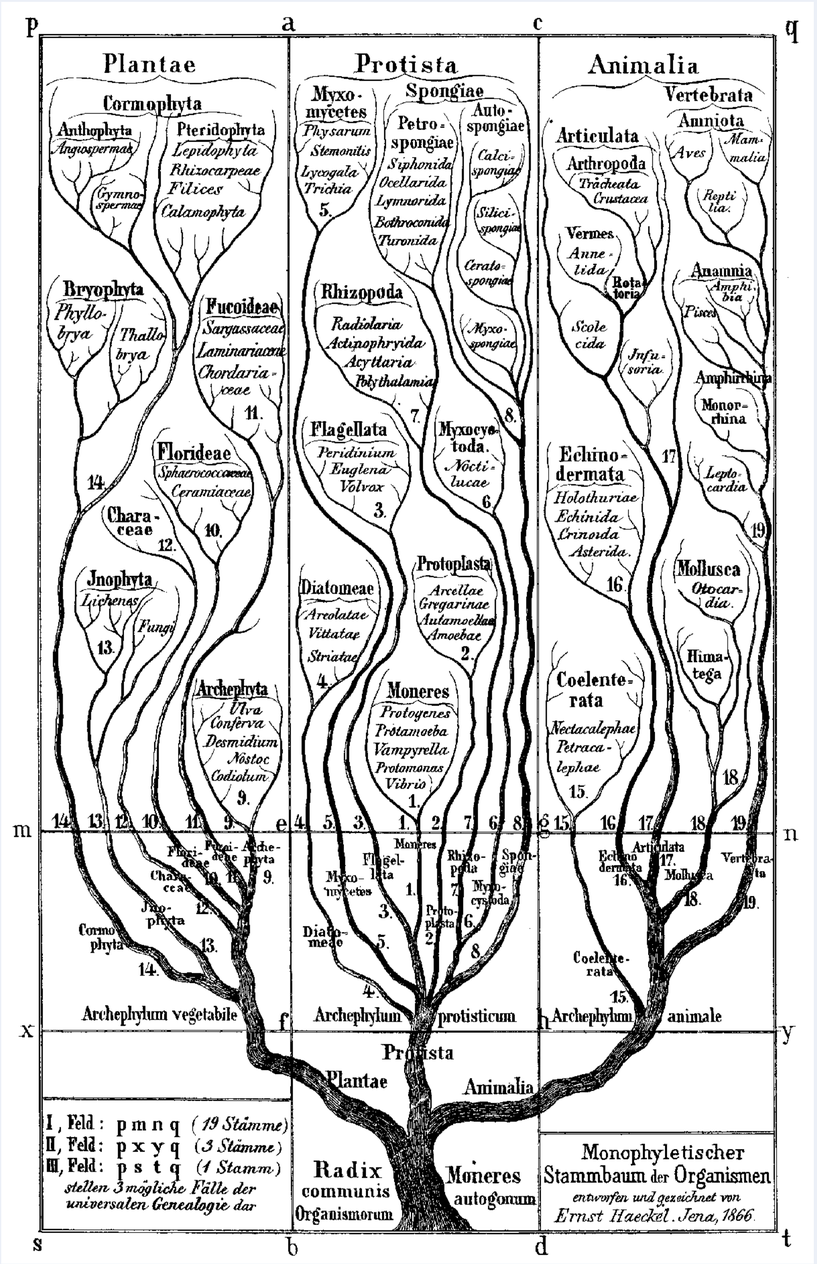
Visual representations of hypotheses about evolutionary relationships



Bacterial evolution. Woese, 1987

Morphology

Molecular sequences



General morphology of organisms, Haeckel, 1866

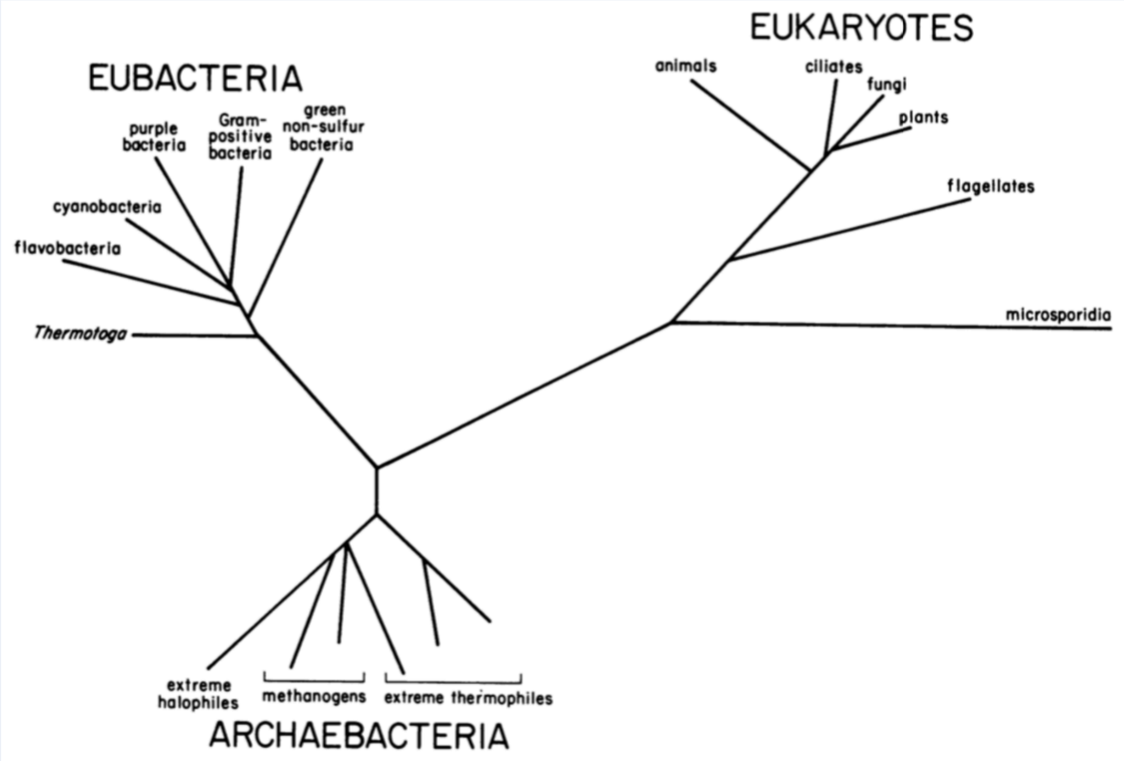
Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

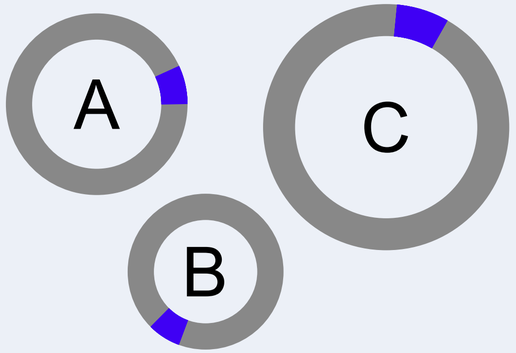
Visual representations of hypotheses about evolutionary relationships



Bacterial evolution. Woese, 1987

Many molecular phylogenetic trees are based on a single gene-type.

1. Identify target gene-type in genomes of interest



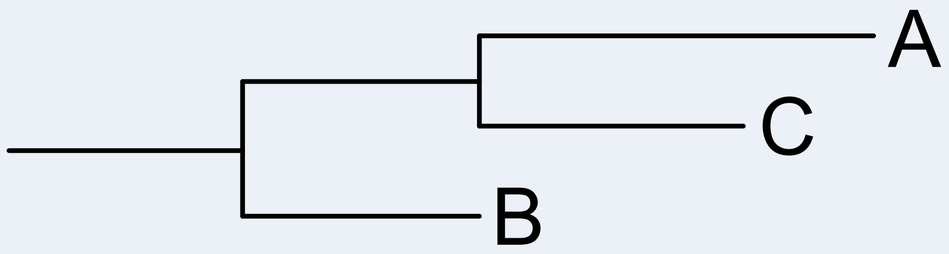
2. Align target genes

Genome **A**'s gene copy `AG-TTAGATC-A`

Genome **B**'s gene copy `-CAT-TGATCAA`

Genome **C**'s gene copy `AG-TTTGATC-A`

3. Infer evolutionary relationships



Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

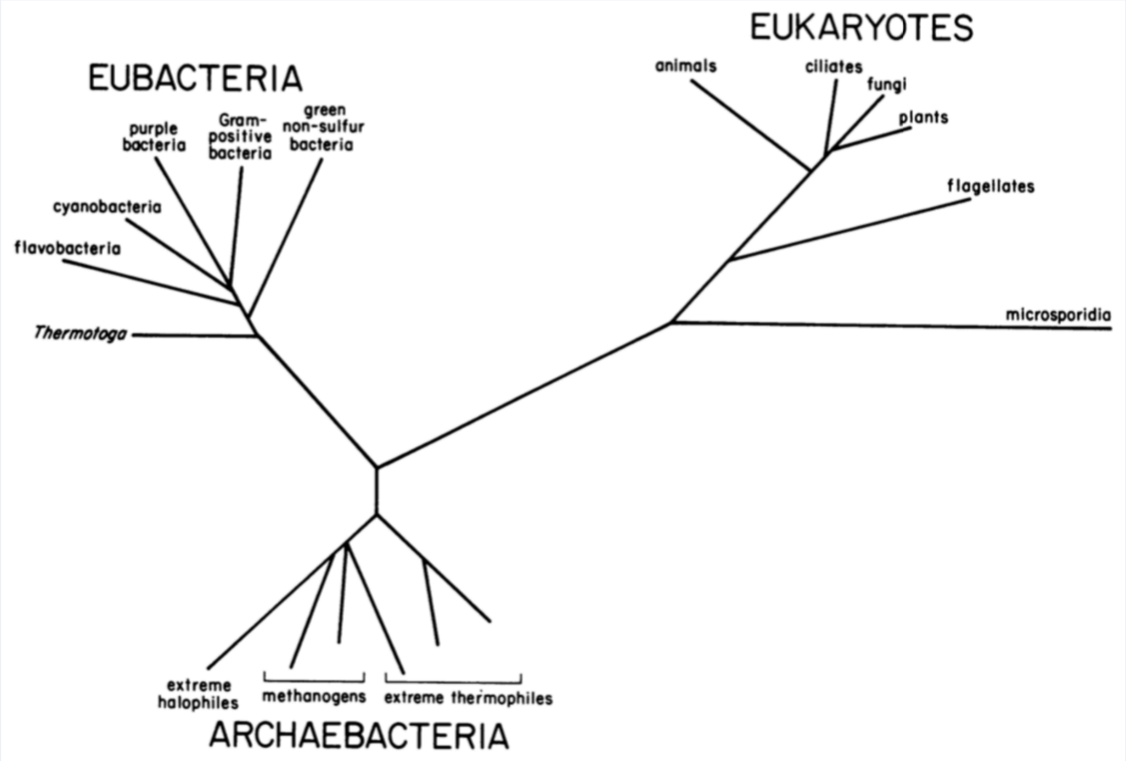
Many molecular phylogenetic trees are based on a single gene-type.

Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships

A phylogenetic tree built from a single gene-type = **A hypothesis about the evolutionary relationships *of those included genes***

(Not the organisms they come from)



Bacterial evolution. Woese, 1987

If we are trying to think about the evolutionary relationships of the *organisms* those genes came from, we are using these genes as proxies standing in for the organisms themselves.

We are assuming that the evolutionary relationships of those individual genes tell us something meaningful about the evolutionary relationships of the organisms they came from.

What makes a gene suitable for this much responsibility??

Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Many molecular phylogenetic trees are based on a single gene-type.

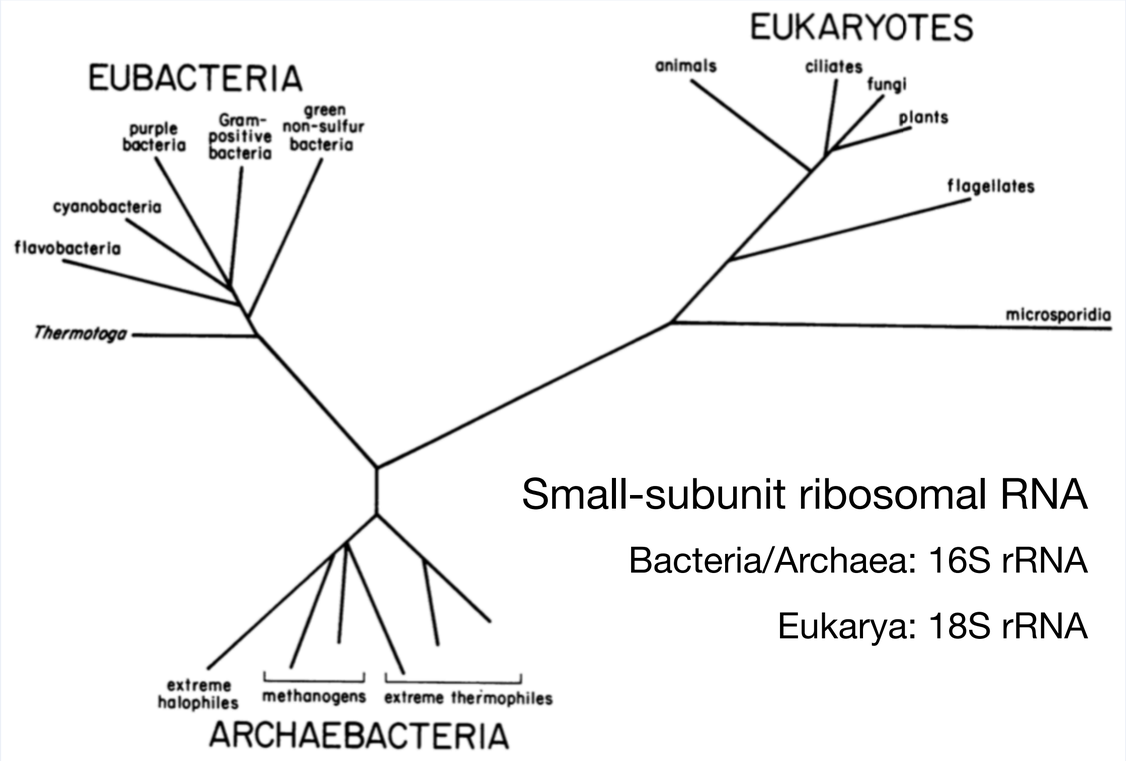
Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships

A phylogenetic tree built from a single gene-type = **A hypothesis about the evolutionary relationships *of those included genes***

(Not the organisms they come from)

Carl Woese pioneered the use of *ribosomal RNA* for this purpose, particularly when attempting to look across all of Life.



Small-subunit ribosomal RNA

Bacteria/Archaea: 16S rRNA

Eukarya: 18S rRNA

Bacterial evolution. Woese, 1987

What are some things we would want in a gene used to represent the evolutionary history of its source organism?

1. Present across all the organisms we want to consider
2. Highly constrained functionally (and in a similar way) across all organisms being considered
   - The more consistently, functionally constrained something is across all target organisms, the less susceptible it is to accruing different evolutionarily selected changes in different organisms
   - We want to measure the “background” accumulation of random mutations (as much as possible)

Ribosome

1. Essential for protein synthesis in all known Bacteria, Archaea, and Eukarya
2. Comprised of many proteins in addition to RNA that all need to interact with each other
   - This helps functionally constrain each of these proteins and the RNA, because a large change in one of these macromolecules might inhibit how the whole unit operates together

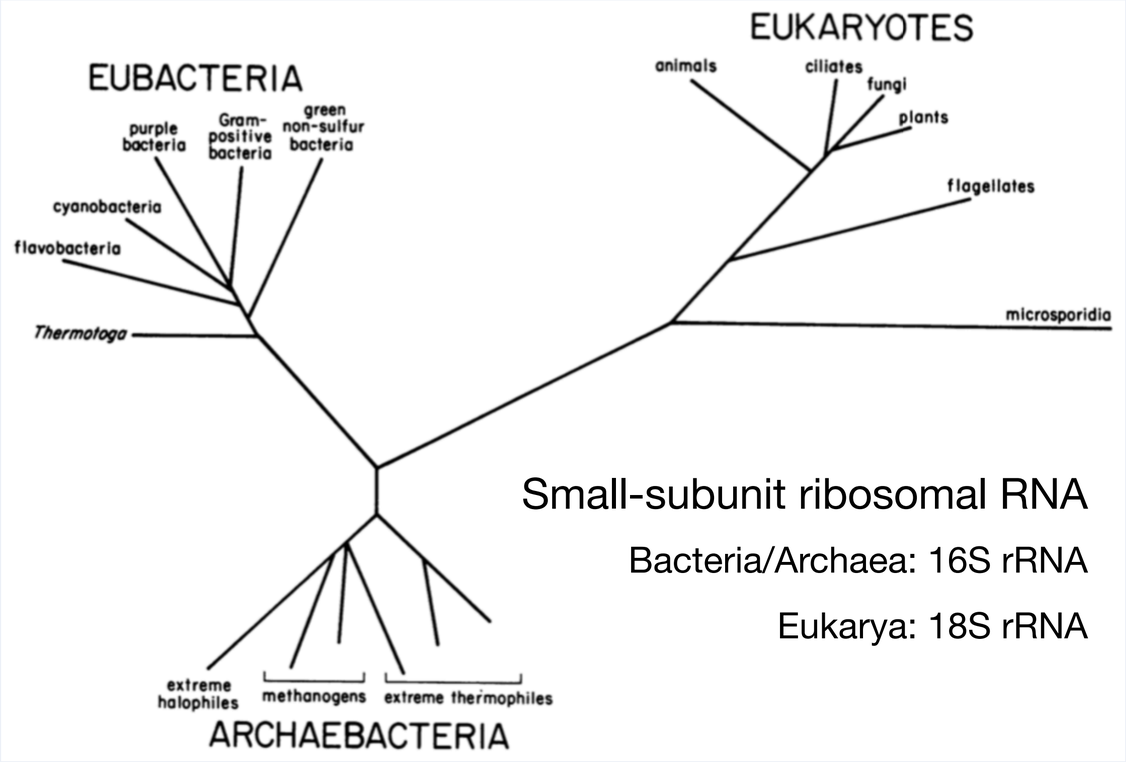
Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

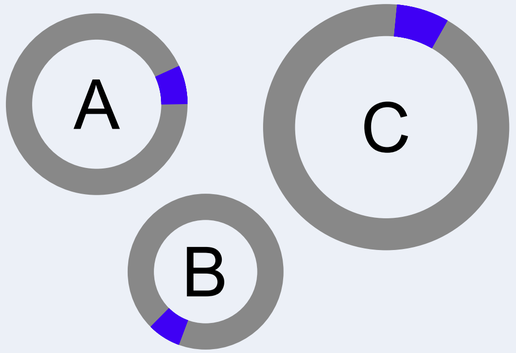
Visual representations of hypotheses about evolutionary relationships



Bacterial evolution. Woese, 1987

Many molecular phylogenetic trees are based on a single gene-type.

1. Identify target gene-type in genomes of interest



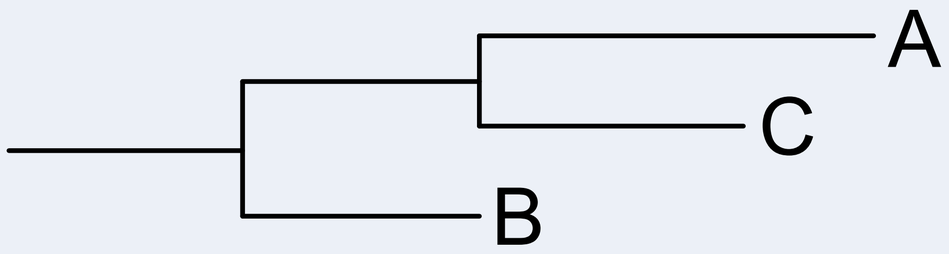
2. Align target genes

Genome **A**'s gene copy `AG-TTAGATC-A`

Genome **B**'s gene copy `-CAT-TGATCAA`

Genome **C**'s gene copy `AG-TTTGATC-A`

3. Infer evolutionary relationships



Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships

Many molecular phylogenetic trees are based on a single gene-type.

System. Appl. Microbiol. 13, 258–262 (1990)

The Case for Relationship of the Flavobacteria and their Relatives to the Green Sulfur Bacteria

C. R. WOESE[1,*], L. MANDELCO[1], D. YANG[1,**], R. GHERNA[2], and M. T. MADIGAN[3]

Summary

Analysis of 16S rRNA sequences suggests, but does not convincingly demonstrate a specific relationship between the eubacterial phylum defined by the flavobacteria and their relatives and that defined by the green sulfur bacteria. Consequently, we have sequenced the 23S rRNA from several representatives of the former group and one of the latter in order to bring more data to bear upon this point. The 23S rRNA data alone strongly suggest a specific relationship between the two phyla, and, together with the 16S rRNA results, provides what we consider now to be a convincing case for this specific relationship.

Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships

Phylogenomics

Under the category of phylogenetics, trying to infer evolutionary relationships at something closer to a genome-level than an individual gene-level phylogeny

Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen[1]

Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

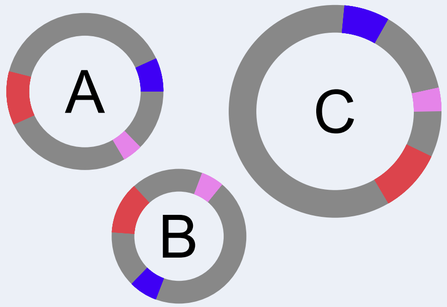
Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships

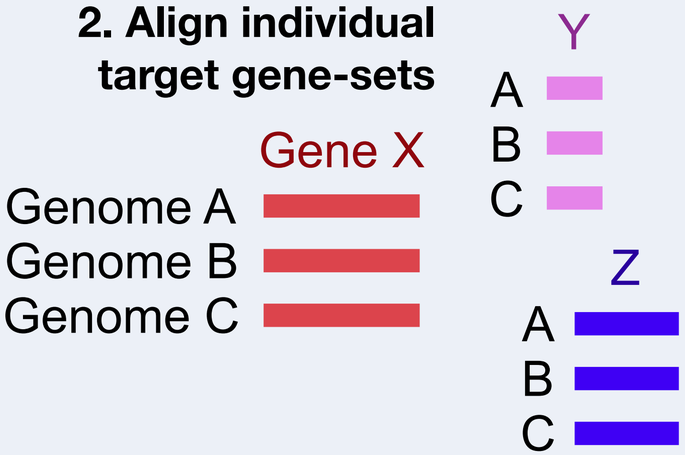
Phylogenomics

Under the category of phylogenetics, trying to infer evolutionary relationships at something closer to a genome-level than an individual gene-level phylogeny

1. Identify target gene-types in genomes of interest

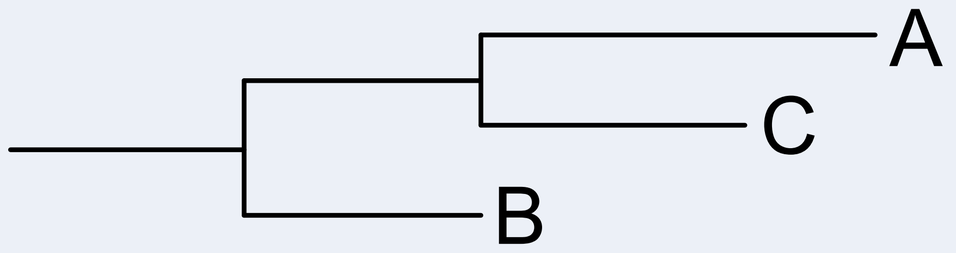


2. Align individual target gene-sets



3. Stick alignments together

4. Infer evolutionary relationships



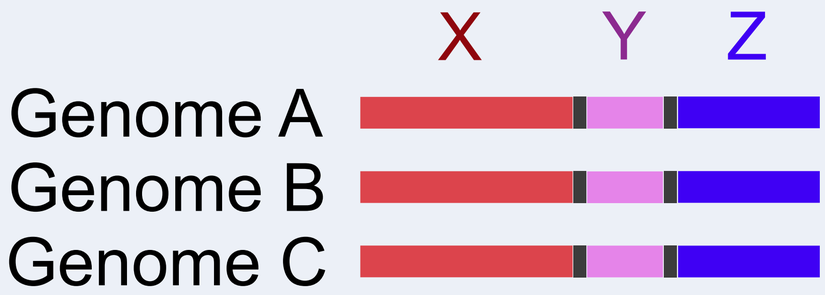
Phylogenetics and phylogenomics

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenetic trees

Visual representations of hypotheses about evolutionary relationships



Phylogenomics

Under the category of phylogenetics, trying to infer evolutionary relationships at something closer to a genome-level than an individual gene-level phylogeny

A phylogenetic tree built from a single gene-type = **A hypothesis about the evolutionary relationships *of those included genes***
(Not the organisms they come from)

A phylogenetic tree built from a concatenated set of genes = **A hypothesis about the evolutionary relationships *of those concatenated genes***
(Not the organisms they come from)

If we are trying to think about the evolutionary relationships of the organisms those genes came from, we are using these sequences we've made by sticking the genes together as proxies standing in for the organisms themselves.

We are assuming that the evolutionary relationships of those sequences tell us something meaningful about the evolutionary relationships of the organisms they came from.

Like before, we have similar guiding principles to help us feel like that assumption is at least reasonable in theory.

Phylogenetics and phylogenomics

Single-copy core genes

Which genes should we use?

Single-copy core genes

*Also just "**s**ingle-**c**opy **g**enes" (SCGs), these are genes that are present in exactly 1 copy in all or most of the organisms we are focusing on.*

What would we want in a gene used to represent the evolutionary history of its host organism?

1. Present across all the organisms we want to consider
   - "core" genes

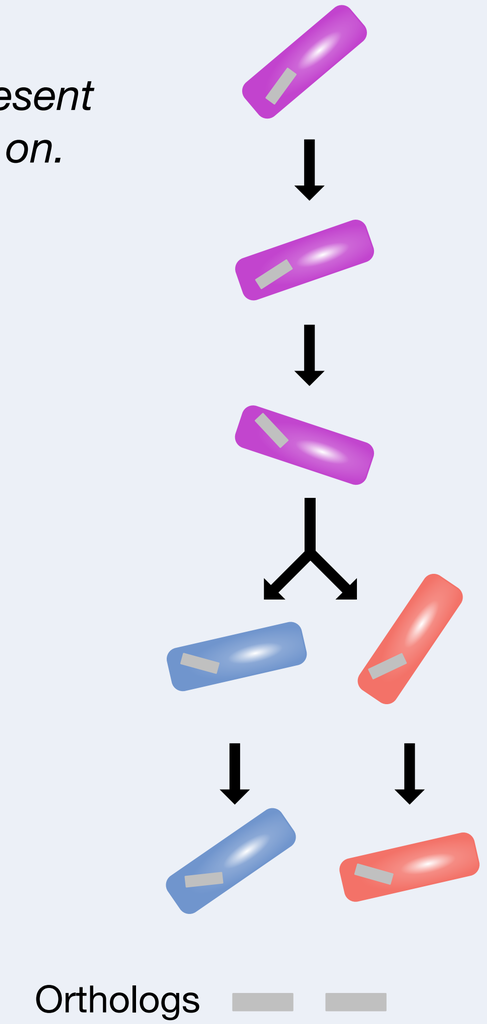
2. Highly constrained functionally across all organisms being considered
   - "single-copy" – when multiple copies of a gene exist within a genome, it becomes likely one of them may be diverging in sequence and function, and now under completely different evolutionary pressures (which we'd want to avoid)
   - We want genes that are "orthologous"

Single-copy core genes

*Also just "**s**ingle-**c**opy **g**enes" (SCGs), these are genes that are present in exactly 1 copy in all or most of the organisms we are focusing on.*

Orthologous genes

Orthologs are versions of the same gene-type in different organisms that have only diverged along with those organisms.



Single-copy core genes

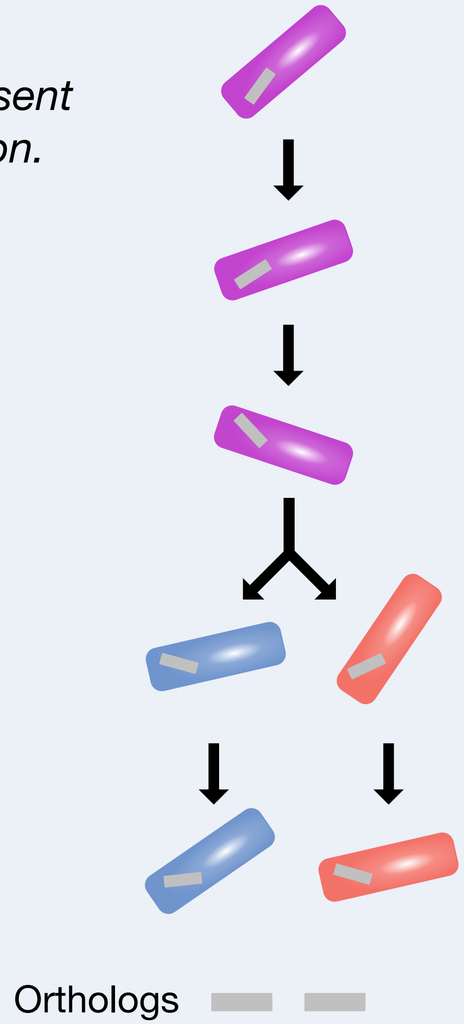
*Also just "**s**ingle-**c**opy **g**enes" (SCGs), these are genes that are present in exactly 1 copy in all or most of the organisms we are focusing on.*

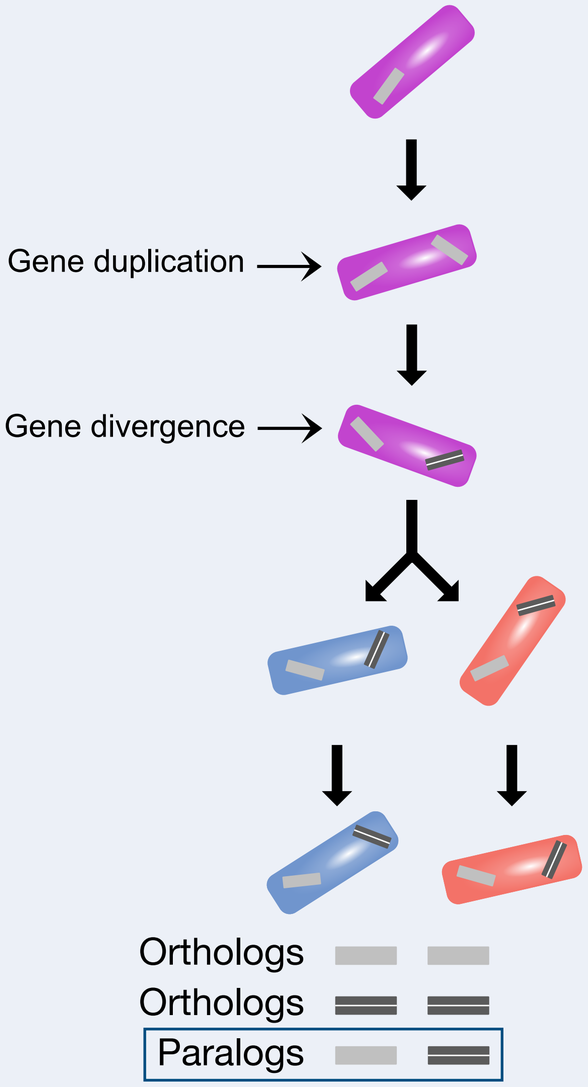
Orthologous genes

Orthologs are versions of the same gene-type in different organisms that have only diverged along with those organisms.

Paralogous genes

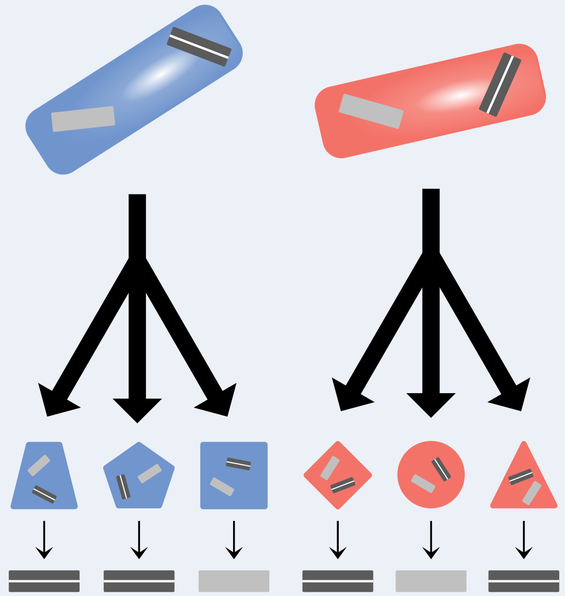
Paralogs are the result of a gene duplication event, and are more likely to be evolving under different evolutionary pressures.





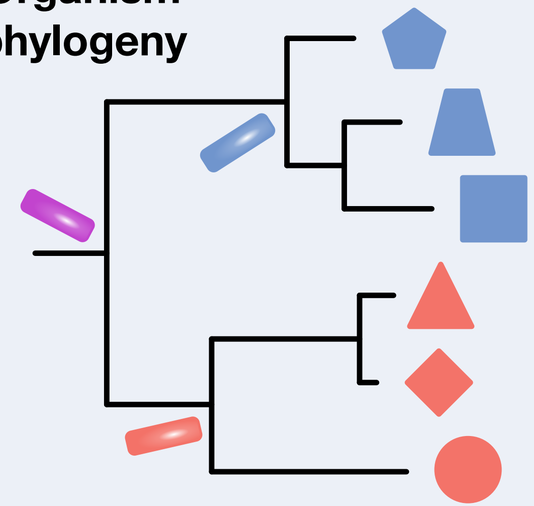
Single-copy core genes

*Also just "**s**ingle-**c**opy **g**enes" (SCGs), these are genes that are present in exactly 1 copy in all or most of the organisms we are focusing on.*

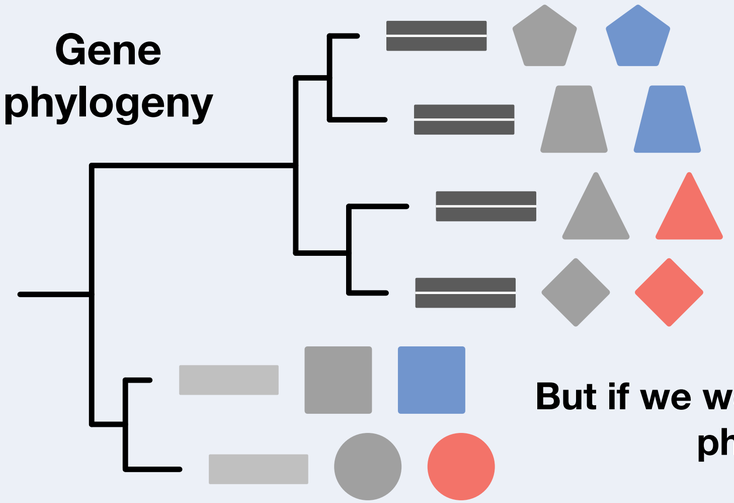


If we were trying to think about the gene-level phylogeny, this would be informative

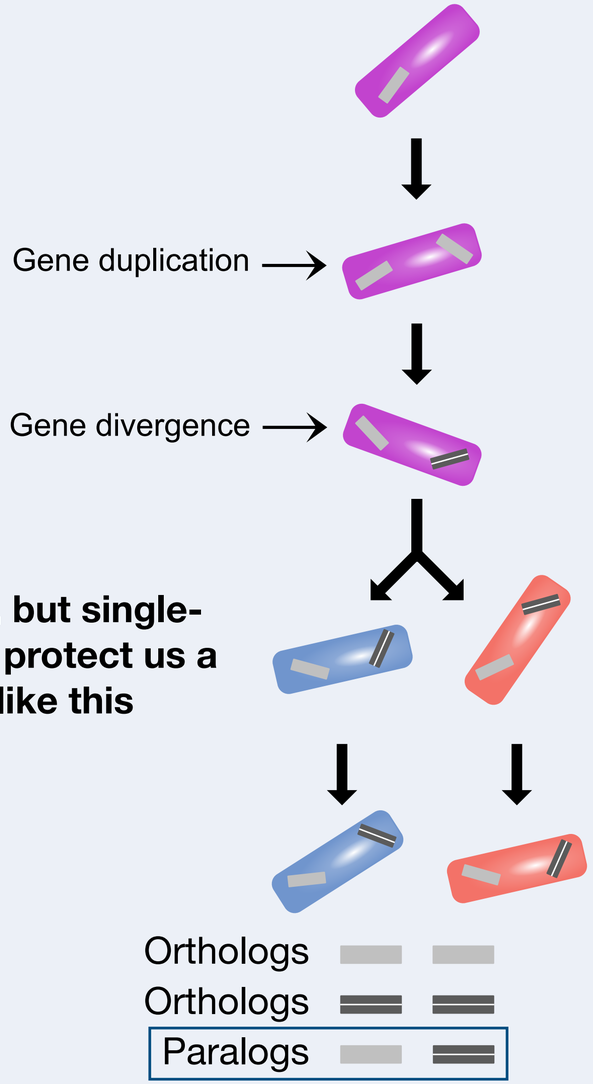
Organism phylogeny



Gene phylogeny



But if we were trying to think about the organism-level phylogeny, we would be misled here



They're not perfect, but single-copy orthologs help protect us a lot from issues like this

Phylogenetics and phylogenomics

Single-copy core genes

Which genes should we use?

Which genes should we use?

The organisms we are focusing on dictate the genes that should be used.

As breadth of diversity being considered  **The number of single-copy core genes shared by all** 

Target organisms	Rough estimate of potentially suitable genes
All 3 domains	~15
Bacteria (domain*)	~70
Cyanobacteria (phylum*)	~250
Marine *Synechococcus* (genus*)	~1,000

Depending on what our purpose is and the span of organisms we are considering, we may want to use a previously generated set of genes that are suitable for a specific clade.

But there may also be times when we want to identify our own single-copy genes that are highly specific to the organisms that we care about.

* These rough numbers are for these specific groups (e.g. Cyanobacteria). They are not necessarily relevant for these taxonomic ranks as a whole (e.g. phylum).

Summary

Phylogenetics

The practice of trying to infer evolutionary relationships between things based on heritable traits or characteristics

Phylogenomics

Under the category of phylogenetics, trying to infer evolutionary relationships at something closer to a genome-level than an individual gene-level phylogeny

Phylogenetic trees

Visual representations of hypothesized evolutionary relationships

Why single-copy core genes?

1. Present across all the organisms we want to consider
   - "core" genes

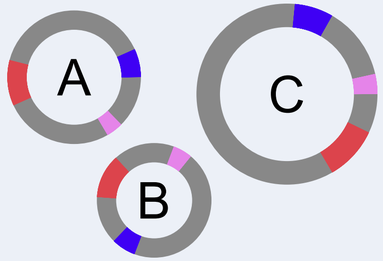
2. Highly constrained functionally across all organisms being considered
   - "single-copy" – we are safer in assuming they are under relatively more similar evolutionary pressures

A molecular phylogenetic tree = **A hypothesis about the evolutionary relationships *of the included sequences***
(Not the organisms they come from)

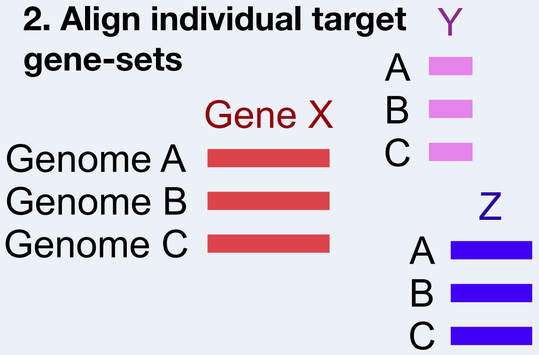
Our genomes of interest determine which genes are suitable

As breadth of diversity being considered increases ↑ **The number of single-copy core genes shared by all decreases** ↓

1. Identify target gene-types in genomes of interest

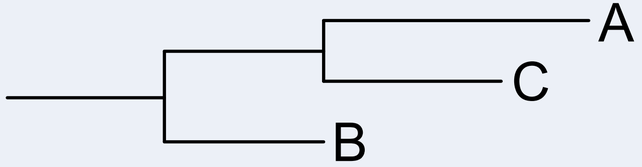


2. Align individual target gene-sets

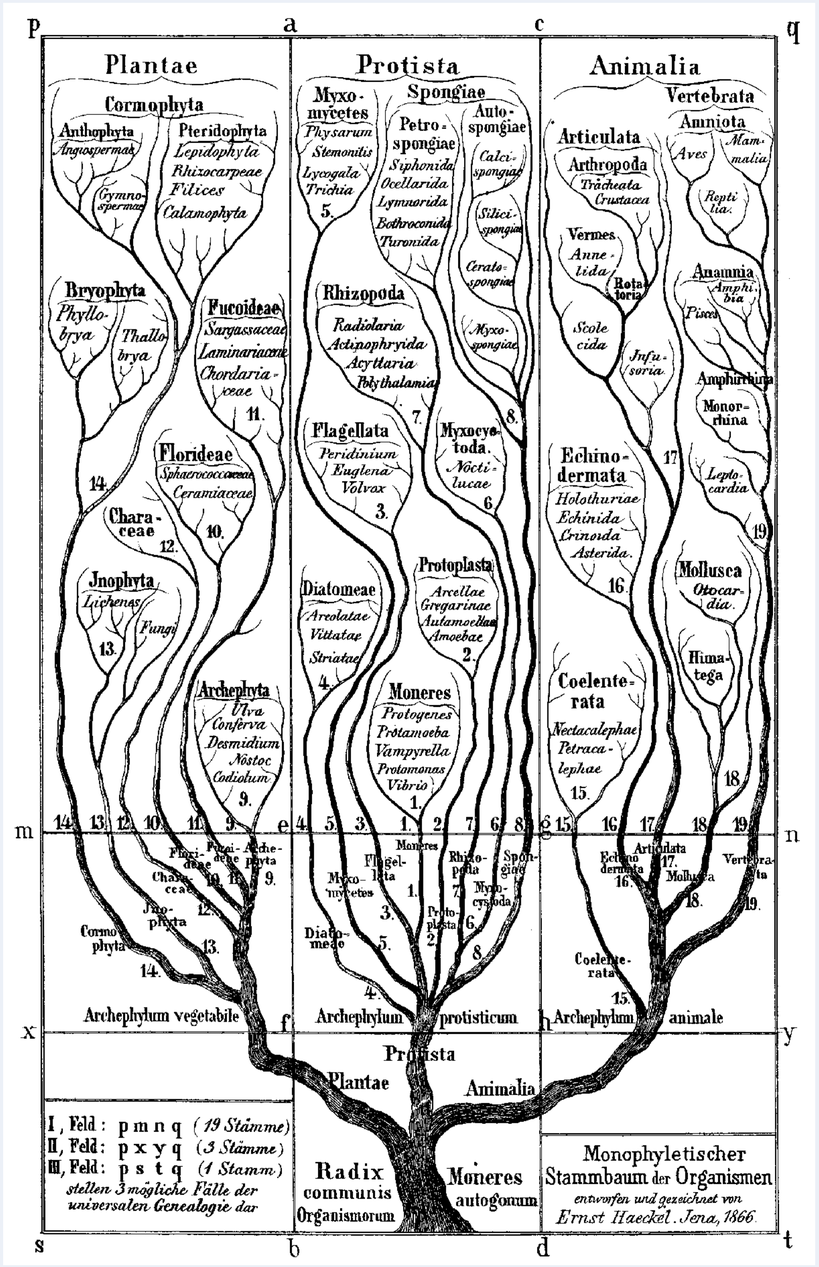


3. Stick alignments together

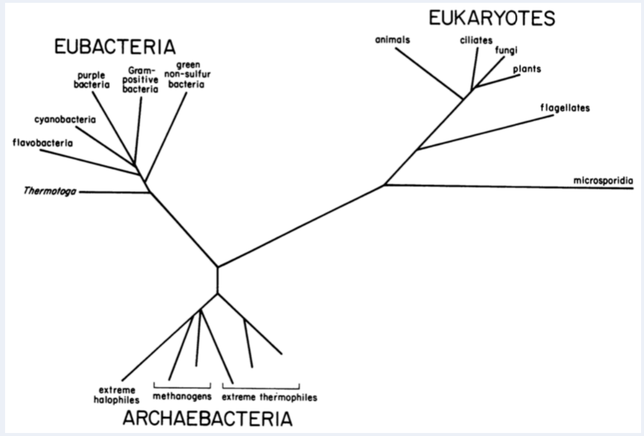
4. Infer evolutionary relationships



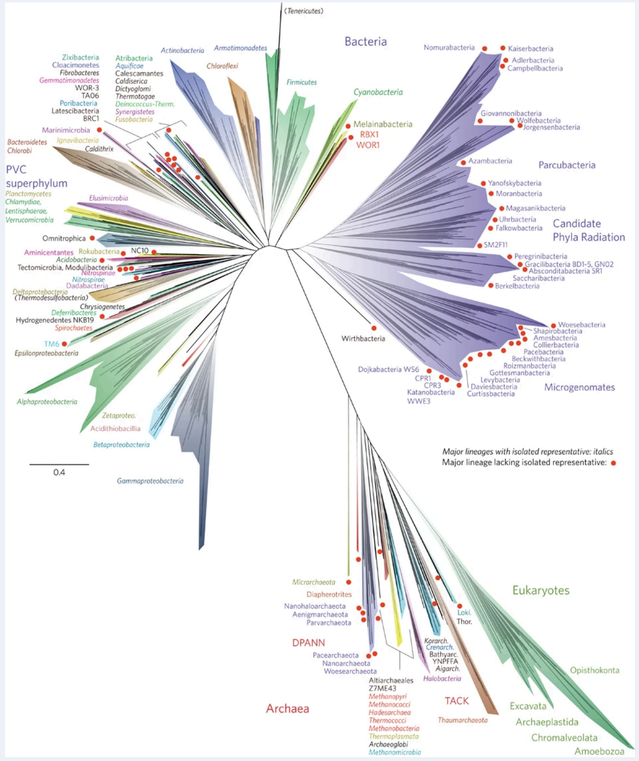
An introduction to phylogenomics



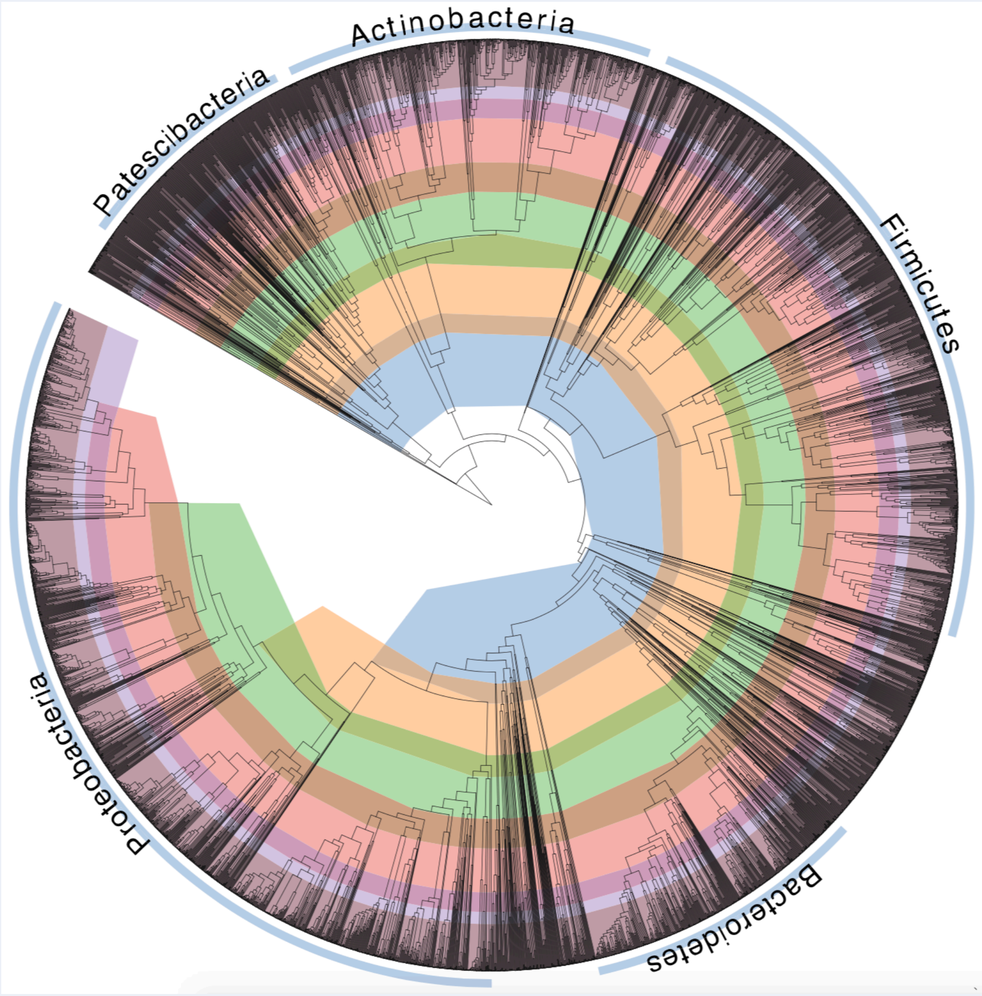
General morphology of organisms, Haeckel, **1866**



Bacterial evolution. Woese, **1987**



A new view of the tree of life. Hug et al., **2017**



A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Parks et al., **2018**

Thanks!

Mike Lee
@AstrobioMike
microbialomics.org

GeneLab
Open Science for Life in Space

Blue Marble Space
Institute of Science